

# Smooth, identifiable supermodels of discrete DAG models with latent variables

**Robin J. Evans**  
Department of Statistics  
University of Oxford  
evans@stats.ox.ac.uk

**Thomas S. Richardson**  
Department of Statistics  
University of Washington  
tsr@stat.washington.edu

November 24, 2015

## Abstract

We provide a parameterization of the discrete nested Markov model, which is a supermodel that approximates DAG models (Bayesian networks) with latent variables. We explicitly evaluate the dimension of such models, show that they are curved exponential families of distributions, and fit them to data. The parameterization avoids the irregularities and unidentifiability of latent variable models. The parameters used are all fully identifiable and causally-interpretable quantities.

## 1 Introduction

Directed acyclic graph (DAG) models, also known as Bayesian networks, are a widely used class of multivariate models in probabilistic reasoning, machine learning and causal inference (Bishop, 2007; Darwiche, 2009; Pearl, 2009). The inclusion of latent variables within Bayesian network models can greatly increase their flexibility, and also account for unobserved confounding; however, latent variable models are typically non-regular, their dimension can be hard to calculate, and they generally do not have fully identifiable parameterizations. In this paper we will present an alternative approach which overcomes these difficulties, and does not require any parametric assumptions to be made about the latent variables.

**Example 1.1.** Suppose we are interested in the relationship between family income during childhood  $X$ , an individual's education level  $E$ , their military service  $M$ , and their later income  $Y$ . We might propose the model shown in Figure 1(a), which includes a hidden variable  $U$  representing motivation or intelligence. Let the four observed variables be binary, but make no assumption about  $U$ .

One can check using Pearl's d-separation criterion (Pearl, 2009) that  $M \perp\!\!\!\perp X \mid E$  under this model, i.e. there is no relationship between military service and family income after controlling for level of education; this places two independent constraints on the distribution  $p(x, e, m, y)$  (one for each level of  $E$ ). In addition, let

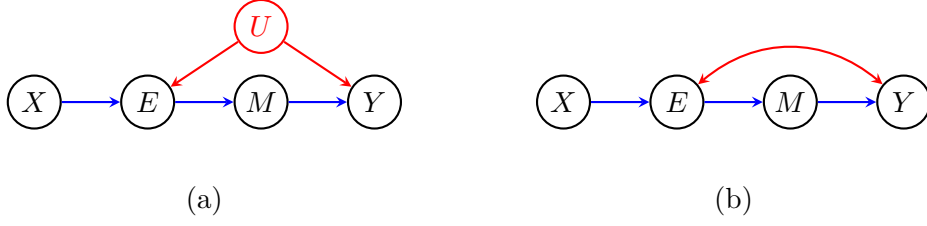


Figure 1: (a) A directed acyclic graph with the latent variable  $U$ ; (b) a (conditional) acyclic directed mixed graph (the *Verma graph*) representing the observed distribution in (a).

$q_{EY}(e, y | x, m) \equiv p(e | x) \cdot p(y | x, m, e)$ ; then the quantity

$$\begin{aligned} q_{EY}(y | x, m) &\equiv \sum_e q_{EY}(e, y | x, m) \\ &= \sum_e p(e | x) \cdot p(y | x, m, e) \end{aligned} \quad (1)$$

does not depend upon  $x$  (Robins, 1986). If the graph is interpreted causally then  $q_{EY}(y | x, m) = p(y | \text{do}(x, m))$ , i.e. it is the distribution of  $Y$  in an experiment that externally sets  $\{X = x, M = m\}$ . Note that generally  $q_{EY}(y | x, m) \neq p(y | x, m)$ .

This restriction on  $q_{EY}$  corresponds to two further independent constraints on  $p$ , one for each level of  $m$ . The set of distributions that satisfy all four constraints is the *nested Markov model* associated with the graph(s) in Figure 1; the number of free parameters is  $15 - 2 - 2 = 11$ .

The distributions in the model all factorize as

$$p(x, e, m, y) = p(x) \cdot p(m | e) \cdot q_{EY}(e, y | x, m),$$

and each of the three factors can be parameterized separately. The model can therefore be described using the following 11 free parameters:

$$\begin{array}{lll} p(x = 0) & p(m = 0 | e) & q_{EY}(e = 0 | x) \\ & q_{EY}(y = 0 | m) & q_{EY}(e = y = 0 | x, m). \end{array}$$

which—if we interpret the model as a causal one—are respectively the quantities

$$\begin{array}{lll} P(X = 0) & P(M = 0 | E = e) & P(E = 0 | X = x) \\ & P(Y = 0 | \text{do}(M = m)) & P(E = 0, Y = 0 | \text{do}(X = x, M = m)). \end{array}$$

The map between positive probability distributions in the nested Markov model and these 11 parameters is smooth and bijective, and the parameters are fully identifiable. It follows that the model is a curved exponential family of distributions, and that it can be fitted using standard numerical methods.

An alternative approach would be to include a latent variable  $U$  explicitly in the model, but this leads to some parameters being unidentifiable. For example, with a binary  $U$  the model implied by Figure 1(a) has 12 parameters. We know that

the true marginal distribution has at most 11 dimensions, so at least one of these 12 parameters is unidentifiable. Even though the model is over parameterized, it may still be that the latent variable model is ‘too small’, in the sense that more distributions over the observed margin can be obtained if  $U$  is allowed to have more than two states. As such, it is undesirable to make specific assumptions about  $U$ ’s state-space. Further, latent variable models are not statistically regular, so standard statistical theory for likelihood ratio tests and asymptotic normality of parameter estimates does not apply (see, e.g. Mond et al., 2003; Drton, 2009).

Models of conditional independence associated with margins of DAG models (we refer to these as ‘ordinary Markov models’) have been studied by Richardson (2003); see also Wermuth (2011). These models were parameterized and shown to be smooth by Evans and Richardson (2014). Other approaches using probit models (Silva and Ghahramani, 2009) and cumulative distribution networks (Huang and Frey, 2008; Silva et al., 2011) are more parsimonious than ordinary Markov models, but impose additional constraints due to their parametric structure.

None of the models mentioned in the previous paragraph can account for constraints of the kind in (1), which were first identified by Robins (1986) and separately by Verma and Pearl (1991). An algorithm for finding such constraints was given by Tian and Pearl (2002b), and developed into a statistical model (the nested Markov model) by Shpitser et al. (2014). Evans (2015) showed that the discrete nested Markov model is the best possible algebraic approximation to DAG models with latent variables, in the sense that the models have the same dimension over the observed variables. In this paper we provide a smooth, statistically regular and fully identifiable parameterization of discrete nested Markov models. As a result, discrete nested Markov models are shown to be curved exponential families of distributions of known dimension. All the parameters we derive are interpretable as straightforward causal quantities.

The remainder of the paper is organized as follows. In Section 2 we introduce Conditional Acyclic Directed Mixed Graphs, the class of graphs we use to represent our models; those models are formally introduced in Section 3. Some graphical theory is given in Section 4, before the main results in Section 5. Section 6 applies the method to data from a panel study.

## 2 Conditional acyclic directed mixed graphs

A directed acyclic graph (DAG) contains vertices representing random variables, and edges (arrows) that imply some structure on the joint probability distribution. A DAG with latent vertices can be transformed into an acyclic directed mixed graph (ADMG) over just its observed vertices via an operation called *latent projection* (Pearl and Verma, 1992). In the simplest case this just involves replacing latent variables with bidirected edges ( $\leftrightarrow$ ), as illustrated by the transformation from Figure 1(a) to (b); the transformed graph represents the marginal distribution over the observed random variables  $X_V$ .

For technical reasons we work with a slightly larger class of graphs, called *conditional* acyclic directed mixed graphs (CADMGs). These have two types of vertex, fixed ( $W$ ) and random ( $V$ ), and are used to represent the structure of a set of distributions for  $X_V$  indexed by  $X_W$ .

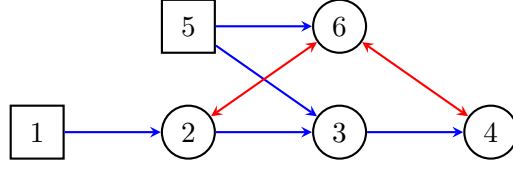


Figure 2: A conditional acyclic directed mixed graph  $\mathcal{L}$ , with random vertices  $V = \{2, 3, 4, 6\}$  and fixed vertices  $W = \{1, 5\}$ .

**Definition 2.1.** A *conditional acyclic directed mixed graph* (CADMG)  $\mathcal{G}$  is a quadruple  $(V, W, \mathcal{E}, \mathcal{B})$ . There are two disjoint sets of vertices: *random*,  $V$ , and *fixed*,  $W$ . The *directed edges*  $\mathcal{E} \subseteq (V \cup W) \times V$  are ordered pairs of vertices; if  $(a, b) \in \mathcal{E}$  we write  $a \rightarrow b$ . Loops  $a \rightarrow a$  and directed cycles  $a \rightarrow \dots \rightarrow a$  are not allowed (hence ‘acyclic’).

The *bidirected edges*,  $\mathcal{B}$ , are unordered pairs of random vertices, and if  $\{a, b\} \in \mathcal{B}$  we write  $a \leftrightarrow b$ .

These graphical concepts are most easily understood by example: see the CADMG in Figure 2. We denote random vertices with round nodes, and fixed vertices with square nodes. CADMGs are not generally simple graphs, because it is possible to have up to two edges between each pair of vertices in  $V$  (one directed and one bidirected). CADMGs are a slight generalization of ADMGs (Richardson, 2003), which correspond to the special case  $W = \emptyset$ . Note that no arrow heads can be adjacent to any fixed vertex: so neither  $a \rightarrow w$  nor  $a \leftrightarrow w$  is allowed for any  $w \in W$ . This reflects the fact that fixed vertices cannot depend on other variables, observed or unobserved, but that random vertices may depend upon fixed ones. Mathematically, fixed nodes play a similar role to the ‘parameter nodes’ used by Dawid (2002).

We make use of the following standard familial terminology for directed graphs.

**Definition 2.2.** If  $a \rightarrow b$  we say that  $a$  is a *parent* of  $b$ , and  $b$  a *child* of  $a$ . The set of parents of  $b$  is denoted  $\text{pa}_{\mathcal{G}}(b)$ . We say that  $w$  is an *ancestor* of  $v$  if either  $v = w$  or there is a sequence of directed edges  $w \rightarrow \dots \rightarrow v$ . The set of ancestors of  $v$  is denoted  $\text{an}_{\mathcal{G}}(v)$ . These definitions are applied disjunctively to sets of vertices so that, for example,  $\text{pa}_{\mathcal{G}}(A) \equiv \bigcup_{a \in A} \text{pa}_{\mathcal{G}}(a)$ . An *ancestral* set is one that contains all its own ancestors:  $\text{an}_{\mathcal{G}}(A) = A$ .

Note that the definitions of parents, children and ancestors do not distinguish between random and fixed vertices. A *random-ancestral* set,  $A' \subseteq V$ , is a set of random vertices such that  $\text{an}_{\mathcal{G}}(A') \subseteq A' \cup W$ ; i.e. all the random ancestors of  $A'$  are contained in  $A'$  itself.

A set of vertices  $B$  is said to be *sterile* if it does not contain any of its parents:  $\text{pa}_{\mathcal{G}}(B) \cap B = \emptyset$ . The sterile vertices in a graph  $\mathcal{G}$  are those vertices without children in  $\mathcal{G}$ . The *sterile subset* of  $C \subseteq V$  is  $\text{sterile}_{\mathcal{G}}(C) \equiv C \setminus \text{pa}_{\mathcal{G}}(C)$ .

**Example 2.3.** Consider the CADMG  $\mathcal{L}$  in Figure 2. The parents of 3 are  $\text{pa}_{\mathcal{L}}(3) = \{2, 5\}$ , and its ancestors are  $\text{an}_{\mathcal{L}}(3) = \{1, 2, 3, 5\}$ ; hence  $\{1, 2, 3, 5\}$  is ancestral, and  $\{2, 3\}$  is random-ancestral. The set  $\{2, 4, 6\}$  is sterile, but  $\{2, 3, 6\}$  is not.

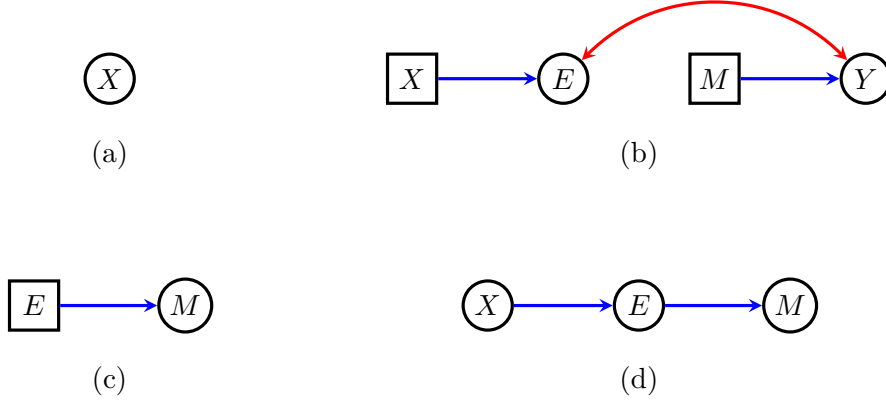


Figure 3: Reachable subgraphs of the graph in Figure 1(b). Graphs in (a), (b) and (c) correspond to factorization into the districts  $\{X\}$ ,  $\{E, Y\}$  and  $\{M\}$  respectively. Graph (d) corresponds to marginalizing the childless node  $Y$ .

For convenience, throughout this paper we will only consider CADMGs in which every fixed vertex has at least one child.

**Definition 2.4.** A set of random vertices  $B \subseteq V$  is *bidirected-connected* if for each  $a, b \in B$  there is a sequence of edges  $a \leftrightarrow \dots \leftrightarrow b$  with all intermediate vertices in  $B$ . A maximal bidirected-connected set is a *district* of the graph  $\mathcal{G}$ . Districts partition the random vertices of a graph; the district containing  $v \in V$  is denoted  $\text{disg}(v)$ .

We draw bidirected edges in red, which makes it easy to identify districts as the maximal sets connected by red edges. In Figure 1(b) for example, there are three districts:  $\{X\}$ ,  $\{M\}$ , and  $\{E, Y\}$ . In Figure 2 there are two:  $\{3\}$  and  $\{2, 4, 6\}$ .

## 2.1 Transformations

We now introduce two operations that transform CADMGs by removing vertices: separating into districts and taking ancestral subgraphs. We will use these transformations to define our Markov property (and thereby our statistical model) in Section 3.

**Definition 2.5.** Let  $\mathcal{G}$  be a CADMG containing a district  $D$ . Define  $\mathfrak{d}_D(\mathcal{G})$  to be the CADMG with random vertices  $D$ , fixed vertices  $\text{pa}_{\mathcal{G}}(D) \setminus D$ , those directed edges from  $\mathcal{G}$  pointing into  $D$ , and bidirected edges whose endpoints are both in  $D$ .

Let  $A$  be a random-ancestral set in  $\mathcal{G}$ . Define  $\mathfrak{m}_A(\mathcal{G})$  to be the graph with random vertices  $A$ , fixed vertices  $\text{pa}_{\mathcal{G}}(A) \setminus A$ , and all edges between these vertices. Note that, since  $A$  is random-ancestral, by definition the vertices in  $\text{pa}_{\mathcal{G}}(A) \setminus A$  are already fixed vertices in  $\mathcal{G}$ .

If a graph  $\mathcal{G}'$  can be obtained from  $\mathcal{G}$  by iteratively applying operations of the form  $\mathfrak{d}$  and  $\mathfrak{m}$ , we say that  $\mathcal{G}'$  is *reachable* from  $\mathcal{G}$ .

**Example 2.6.** The graph in Figure 1(b) contains the districts  $\{X\}$ ,  $\{E, Y\}$  and  $\{M\}$ . The corresponding graphs  $\mathfrak{d}_D(\mathcal{G})$  are given in Figure 3(a), (b) and (c) re-

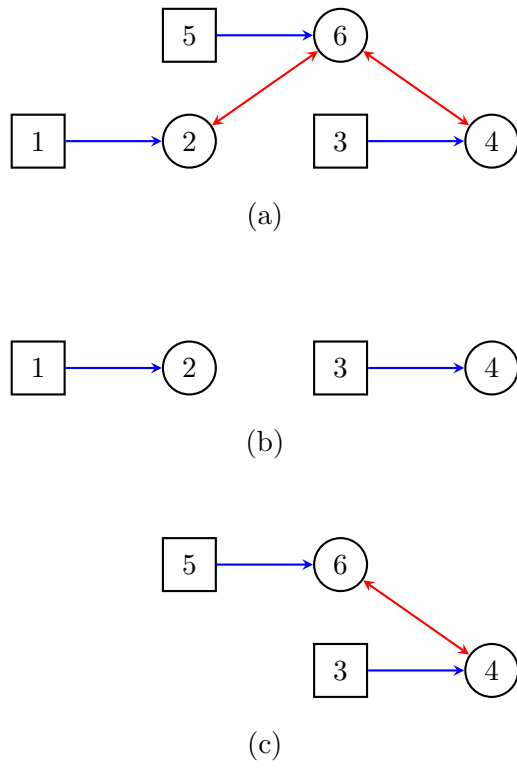


Figure 4: Three CADMGs reachable from the graph in Figure 2.

spectively. The set  $\{X, E, M\}$  is ancestral, and the graph  $\mathbf{m}_{\{X, E, M\}}(\mathcal{G})$  is shown in Figure 3(d).

**Example 2.7.** The graph in Figure 2 contains the district  $\{2, 4, 6\}$ , and  $\mathfrak{d}_{\{2, 4, 6\}}(\mathcal{L})$  gives us the graph in Figure 4(a). The sets  $\{2, 4\}$  and  $\{4, 6\}$  are both random-ancestral in  $\mathfrak{d}_{\{2, 4, 6\}}(\mathcal{L})$ , so we can apply either  $\mathbf{m}_{\{2, 4\}}$  or  $\mathbf{m}_{\{4, 6\}}$  to obtain the CAD-MGs in Figures 4(b) and (c) respectively.

As we will see in the next section, both of these graphical operations correspond to an operation on a probability distribution we associate with the graph:  $\mathbf{m}_A$  to marginalization, and  $\mathfrak{d}_D$  to a factorization. The ‘fixing’ operation described in Shpitser et al. (2014) unifies and generalizes  $\mathbf{m}$  and  $\mathfrak{d}$ , but the statistical model we will describe is ultimately the same. For the purposes of defining a parameterization it is more convenient to use the formulation given here.

Note that if we start with a graph  $\mathcal{G}$  in which all the fixed vertices  $w \in W$  have at least one child, then this is also true of the graph obtained after applying either  $\mathbf{m}_A$  or  $\mathfrak{d}_D$ . If a graph with random vertices  $C$  is reachable from  $\mathcal{G}$  then we will denote it  $\mathcal{G}[C]$ ; the next lemma shows that this is well defined.

It is important to note that sets may become districts or random ancestral sets after several iterations of  $\mathbf{m}$  and  $\mathfrak{d}$ . For example,  $\{2, 6\}$  is not random-ancestral in  $\mathcal{L}$ , but it is in  $\mathfrak{d}_{\{2, 4, 6\}}(\mathcal{L})$ . Conversely  $\{4\}$  is not a district in  $\mathfrak{d}_{\{2, 4, 6\}}(\mathcal{L})$ , but it is in  $\mathbf{m}_{\{2, 4\}}(\mathfrak{d}_{\{2, 4, 6\}}(\mathcal{L}))$ ; see Figure 4(b).

**Lemma 2.8.** *Suppose that the graph  $\mathcal{G}'$  is reachable from  $\mathcal{G}$  and has random vertices  $C$ . Then  $\mathcal{G}'$  is the unique CADMG with random vertices  $C$ , fixed vertices  $\text{pa}_{\mathcal{G}}(C) \setminus C$ , those bidirected edges in  $\mathcal{G}$  with both endpoints in  $C$ , and those directed edges that are directed from  $C \cup \text{pa}_{\mathcal{G}}(C)$  to  $C$ .*

*Proof.* Since we assume all fixed vertices have at least one child, then  $\mathcal{G} = \mathcal{G}[V]$ . In addition it is clear from the definitions of  $\mathfrak{d}$  and  $\mathbf{m}$  that precisely the edges and fixed vertices mentioned are preserved at each step.  $\square$

An equivalent formulation to the Lemma would be to say that  $\mathcal{G}[C]$  has precisely the edges whose arrowheads are all in  $C$ .

### 3 Nested Markov Property

Graphical models relate the structure of a graph to a collection of probability distributions over a set of random variables. We will work with the nested Markov property, which relates a (C)ADMG and each of its reachable subgraphs to a collection of (conditional) probability distributions.

Suppose we are interested in random variables  $X_v$  taking values in a finite discrete set  $\mathfrak{X}_v$ . For a set of vertices  $C$  denote  $\mathfrak{X}_C \equiv \times_{v \in C} \mathfrak{X}_v$ . A *probability kernel for  $V$  given  $W$*  (or simply a kernel) is a function  $p_{V|W} : \mathfrak{X}_V \times \mathfrak{X}_W \rightarrow [0, 1]$  such that for each  $x_W \in \mathfrak{X}_W$ ,

$$\sum_{x_V \in \mathfrak{X}_V} p_{V|W}(x_V | x_W) = 1.$$

In other words, a kernel behaves like a conditional probability distribution for  $X_V$  given  $X_W$ . We use the word ‘kernel’ to emphasize that some of the conditional distributions we obtain are not equal to the usual conditional distribution obtained from elementary definitions, but instead correspond to certain interventional quantities.

In what follows,  $\dot{\cup}$  is used to denote a union of disjoint sets.

**Definition 3.1.** Let  $p_{V|W}$  be a kernel, and let  $A \dot{\cup} B \dot{\cup} C = V$ . The *marginal kernel* over  $A, B | W$  is

$$p_{AB|W}(x_A, x_B | x_W) = \sum_{x_C} p_{V|W}(x_V | x_W).$$

It is easy to check that  $p_{AB|W}$  is also a kernel. The *conditional kernel* of  $A|B, W$  is any kernel  $p_{A|BW}$  satisfying

$$p_{A|BW}(x_A | x_B, x_W) \cdot p_{B|W}(x_B | x_W) = p_{AB|W}(x_A, x_B | x_W).$$

This is uniquely defined precisely for  $x_B, x_W$  such that  $p_{B|W}(x_B | x_W) > 0$ .

**Remark 3.2.** Note that, for convenience, if some of the fixed variables  $W^* \subseteq W$  in a kernel  $p_{V|W}$  are entirely irrelevant, (i.e. if the functions  $p_{V|W}(\cdot | \cdot, y_{W^*})$  are identical for all  $y_{W^*} \in \mathfrak{X}_{W^*}$ ) we will describe it interchangeably as a kernel of  $V$  given  $W$ , and as a kernel of  $V$  given  $W \setminus W^*$ , since in this case these objects are isomorphic:  $p_{V|W} = p_{V|W \setminus W^*}$ .

We are now in a position to define the nested model. The definition is recursive, and works by reference to the model applied iteratively to smaller and smaller graphs. The model is introduced in Shpitser et al. (2014), and is based on the constraint finding algorithm of Tian and Pearl (2002b), which follows a similar recursive structure.

**Definition 3.3.** Let  $\mathcal{G}$  be a CADMG and  $p_{V|W}$  a probability kernel. Say that  $p_{V|W}$  *recursively factorizes* according to  $\mathcal{G}$ , and write  $p_{V|W} \in \mathcal{M}_{rf}(\mathcal{G})$  if either  $|V| = 1$ , or both:

- (i) if  $\mathcal{G}$  has districts  $D_1, \dots, D_k$ ,  $k \geq 1$  then  $p_{V|W} = \prod_i r_i$  where each  $r_i$  recursively factorizes according to  $\mathcal{G}[D_i] = \mathfrak{d}_{D_i}(\mathcal{G})$ ; and
- (ii) for each ancestral set  $A$ , the marginal distribution

$$p_{A \cap V | A \cap W}(x_{A \cap V} | x_{A \cap W}) \equiv \sum_{x_{V \setminus A}} p_{V|W}(x_V | x_W)$$

recursively factorizes according to  $\mathcal{G}[V \cap A] = \mathfrak{m}_{V \cap A}(\mathcal{G})$ .

**Remark 3.4.** It is important to note that, in terms of the factors  $r_i$  whose existence is implied by condition (i), the definition of recursive factorization ‘starts from scratch’ each time we perform the recursion. For example, we make no claim (yet) about the connection between a factor  $r_i$  obtained from (i) and any such factors which arise after first applying (ii) and then later (i): see Example 3.5.

In the base case  $V = \{v\}$  the definition places no restriction on the distribution of  $X_v$  given  $X_W$ . Conditions (i) and (ii) are both satisfied by a distribution obtained



from a directed acyclic graph model with latent variables (Pearl and Verma, 1992; Pearl, 2009). The models defined by the Markov properties for ADMGs introduced by Richardson (2003) and parameterized by Evans and Richardson (2013, 2014) can be defined by replacing (i) with the weaker requirement:

- (i')  $\mathcal{G}$  has districts  $D_1, \dots, D_k, k \geq 1$ , and  $p_{V|W} = \prod_i r_i$  where each  $r_i$  is a kernel for  $D_i$  given  $\text{pa}_{\mathcal{G}}(D_i) \setminus D_i$ .

In other words, although the distribution must satisfy the ancestrality condition and then factorize, no further conditions are imposed on those factors: they are not required to obey any structure implied by the graph  $\mathcal{G}[D_i]$ . This leads to a model defined entirely by conditional independence relations on the original joint distribution  $p_{V|W}$ .

As a consequence of this, the m-separation criterion for ordinary Markov models (as well as the other Markov properties described by Richardson, 2003) can be applied correctly to nested models; however, it does not completely describe the model.

**Example 3.5.** Consider the CADMG in Figure 1(b). Criterion (i) of recursive factorization requires that

$$p(x, e, m, y) = r_X(x) \cdot r_{EY}(e, y | x, m) \cdot r_M(m | e)$$

for distributions  $r_X$ ,  $r_{EY}$  and  $r_M$  which recursively factorize according to the graphs in Figure 3(a), (b) and (c) respectively.

On the other hand, if we apply condition (ii) to the childless node  $Y$  we see that the margin  $p(x, e, m)$  must satisfy recursive factorization with respect to the DAG in Figure 3(d), so

$$p(x, m, e) = \tilde{r}_X(x) \cdot \tilde{r}_E(e | x) \cdot \tilde{r}_M(m | e)$$

for some kernels  $\tilde{r}_X$ ,  $\tilde{r}_M$  and  $\tilde{r}_E$ . This factorization implies the conditional independence  $X \perp\!\!\!\perp M | E$ , which can also be deduced using m-separation. We add a tilde to the kernels to emphasise that the definition starts afresh at each iteration, and makes no claim of any relationship between this factorization and the factorization of  $p = r_X r_{EY} r_M$ . However, it is not hard to verify that in fact

$$\begin{aligned} r_X(x) &= \tilde{r}_X(x) = p(x), \\ r_M(m | e) &= \tilde{r}_M(m | e) = p(m | e), \\ \sum_y r_{EY}(e, y | x, m) &= \tilde{r}_E(e | x) = p(e | x). \end{aligned}$$

Note that

$$\begin{aligned} r_{EY}(e, y | x, m) &= p(e | x) \cdot p(y | x, m, e) \\ &\neq p(e | x, m) \cdot p(y | x, m, e) \\ &= p(e, y | x, m), \end{aligned}$$

and so  $r_{EY}$  is *not* the ordinary conditional distribution of  $E, Y$  given  $X, M$ .

### 3.1 Properties of the Recursive Kernels

The indirectness of Definition 3.3 means that the nature of the kernels  $r_i$  and the connections between them are not generally clear. Here we provide some properties of these kernels, showing that they are products of conditional distributions derived from the original joint distribution, and that they are uniquely defined up to versions of those conditional distributions.

A *topological ordering* of the random vertices of a CADMG is a total ordering  $<$  on  $V$  such that every vertex precedes its children. We denote by  $\text{pre}_{<}(v)$  the set of (random) vertices which precede  $v$  under  $<$ .

The following proposition shows that the factors in the definition of recursive factorization are unique up to versions of conditional distributions.

**Proposition 3.6.** *Let  $\mathcal{G}$  be a CADMG with districts  $D_1, \dots, D_k$ , and let  $<$  be any topological ordering of  $V$ . Let  $p_{V|W} = \prod_i r_i$ , where each  $r_i$  recursively factorizes with respect to  $\mathcal{G}[D_i]$ . Then*

$$r_i(x_{D_i} | x_{\text{pa}_{\mathcal{G}}(D_i) \setminus D_i}) = \prod_{v \in D_i} p_{v | \text{pre}_{<}(v) \cup W}(x_v | x_{\text{pre}_{<}(v)}, x_W), \quad (2)$$

where  $p_{v | \text{pre}_{<}(v) \cup W}$  is any  $p_{V|W}$ -version of the conditional distribution of  $X_v | X_{\text{pre}_{<}(v)}, X_W$ .

**Remark 3.7.** The equation in (2) is an instance of the *g-formula* of Robins (1986).

*Proof.* For the purposes of induction we generalize the result slightly to allow  $D_i$  to be collections of several districts. Let  $E_i \equiv \text{pa}_{\mathcal{G}}(D_i) \setminus D_i$ . We proceed by induction on  $|V|$ : if  $|V| \leq 1$  there is nothing to show. Otherwise, let  $w \in D_k$  be the last vertex in the ordering  $<$ , so that  $x_w$  only appears as a variable in the factor  $r_k$ . Then

$$\begin{aligned} p_{V \setminus w | W}(x_{V \setminus w} | x_W) &\equiv \sum_{x_w} p_{V|W}(x_V | x_W) \\ &= \sum_{x_w} \prod_{i=1}^k r_i(x_{D_i} | x_{E_i}) \\ &= \left( \sum_{x_w} r_k(x_{D_k} | x_{E_k}) \right) \prod_{i=1}^{k-1} r_i(x_{D_i} | x_{E_i}) \\ &= \tilde{r}_k(x_{D_k \setminus w} | x_{E_k}) \prod_{i=1}^{k-1} r_i(x_{D_i} | x_{E_i}) \end{aligned}$$

where, by property 1 of recursive factorization, the kernel  $\tilde{r}_k$  recursively factorizes with respect to the graph  $\mathcal{G}[D_k \setminus \{w\}]$ . Similarly all the factors  $r_i$  for  $i = 1, \dots, k-1$  recursively factorize with respect to  $\mathcal{G}[D_i]$ , so by the induction hypothesis each such  $r_i$  is of the required form (2), and

$$\tilde{r}_k(x_{D_k \setminus \{w\}} | x_{E_k}) = \prod_{v \in D_k \setminus \{w\}} p_{v | \text{pre}_{<}(v) \cup W}(x_v | x_{\text{pre}_{<}(v)}, x_W).$$

But then

$$\prod_i r_i = p_{V|W} = p_{w|VW \setminus \{w\}} \cdot p_{V \setminus \{w\} | W} = p_{w|VW \setminus \{w\}} \cdot \tilde{r}_k \cdot \prod_{i=1}^{k-1} r_i; \quad (3)$$

therefore

$$r_k(x_{D_k} | x_{E_k}, x_W) = p_{w|VW \setminus \{w\}}(x_w | x_{V \setminus \{v\}}, x_W) \cdot \tilde{r}_k$$

and  $p_{w|VW \setminus \{w\}}$  satisfies (3) if and only if it is a version of the relevant conditional distribution, as required.  $\square$

The next result shows that the positivity of  $p_{V|W}$  is preserved in any derived kernels.

**Lemma 3.8.** *Let  $p_{V|W}(x_V | x_W)$  be a probability distribution,  $<$  some total ordering on  $V$ , and let  $A \subseteq V$  and  $B \equiv W \cup \text{pre}_{<}(A) \setminus A$ . Define*

$$r_{A|B}(x_A | x_B) \equiv \prod_{v \in A} p_{v|\text{pre}_{<}(v), W}(x_v | x_{\text{pre}_{<}(v)}, x_W),$$

for some versions  $p_{v|\text{pre}_{<}(v), W}$  of the conditional distributions of  $X_v | X_W, X_{\text{pre}_{<}(v)}$ . Then:

(a)  $r_{A|B}$  is a kernel for  $X_A | X_B$ ;

(b) for any  $T \subseteq V$ ,  $x_T \in \mathfrak{X}_T$  and  $x_W \in \mathfrak{X}_W$ , if  $p_{T|W}(x_T | x_W) > 0$  then

$$r_{T \cap A | B}(x_{T \cap A} | x_B) \equiv \sum_{x_{A \setminus T}} r_{A|B}(x_A | x_B) > 0$$

and all versions of  $r_{T \cap A | B}(x_{T \cap A} | x_B)$  are the same;

(c) if  $p_{T|W}(x_T | x_W) = 0$  then there exists  $t \in T$  such that (every version of)

$$p_{t|\text{pre}_{<}(t), W}(x_t | x_{\text{pre}_{<}(t)}, x_W) = 0.$$

*Proof.* (a) Clearly  $r_{A|B} \geq 0$  since it is a product of conditional distributions, which are themselves non-negative. In addition, by summing the expression above in reverse order of  $<$  it is easy to see that  $\sum_{x_A} r_{A|B}(x_A | x_B) = 1$  for any  $x_B \in \mathfrak{X}_B$ . Hence  $r_{A|B}$  is a kernel.

For (b) note that if  $p_{T|W}(x_T | x_W) > 0$  then there exists some  $x_{V \setminus T} \in \mathfrak{X}_{V \setminus T}$  such that  $p_{V|W}(x_V | x_W) > 0$ . Then

$$\begin{aligned} p_{V|W}(x_V | x_W) &= \prod_{v \in V} p_{v|\text{pre}_{<}(v), W}(x_v | x_{\text{pre}_{<}(v)}, x_W) \\ &= r_{A|B}(x_A | x_B) \prod_{v \in V \setminus A} p_{v|\text{pre}_{<}(v), W}(x_v | x_{\text{pre}_{<}(v)}, x_W), \end{aligned}$$

so if the left hand side is positive then so is  $r_{A|B}(x_A | x_B) > 0$ . Since all the events in this expression have positive  $p_{V|W}$  probability, all versions of each conditional probability are equal.

Lastly, if  $p_{T|W}(x_T | x_W) = 0$  then clearly some factor of

$$0 = p_{T|W}(x_T | x_W) = \prod_{t \in T} p_{t|\text{pre}_{<}(t), W}(x_t | x_{\text{pre}_{<}(t)}, x_W)$$

is also zero. Pick the  $<$ -minimal  $t$  such that this holds, so that  $p_{\text{pre}_{<}(t) | W}(x_{\text{pre}_{<}(t)} | x_W) > 0$ . Then (c) holds.  $\square$

A corollary of this lemma is that, if  $p_{V|W}$  is strictly positive, the kernels  $r_i$  derived from it by application of Definition 3.3 are uniquely defined.

[Check the below is used.]

**Corollary 3.9.** *Let  $p_{V|W}$  be a strictly positive kernel that recursively factorizes according to  $\mathcal{M}_{\text{rf}}(\mathcal{G})$ . Then any kernel derived from  $p_{V|W}$  by repeated applications of Definition 3.3 is uniquely defined.*

*Proof.* Clearly applying (ii) is always unique, since it only involves summing. By Proposition 3.6, application of (i) is a factorization into univariate conditional distributions, each of which is uniquely defined when the joint distribution is positive. In addition, by Lemma 3.8 each such conditional distribution is also strictly positive, so following the recursion with each unique factor gives the result.  $\square$

## 4 Intrinsic Sets and Partitions

In this section we provide the necessary theory to link the graphical notions of Section 3 to the parameterization in Section 5. The parameterization uses factorizations of the distribution into pieces which correspond to special subsets of vertices in the graph; these subsets are themselves derived from the idea of the ‘reachable’ sets already introduced.

**Definition 4.1.** Let  $\mathcal{G}$  be a CADMG. A non-empty set  $S$  of random vertices is *intrinsic* if it is bidirected-connected and the graph  $\mathcal{G}[S]$  is reachable from  $\mathcal{G}$ . Denote the collection of all intrinsic sets in  $\mathcal{G}$  by  $\mathcal{I}(\mathcal{G})$ .

For each intrinsic set  $S \in \mathcal{I}(\mathcal{G})$ , define the associated *recursive head* by  $\text{rh}_{\mathcal{G}}(S) = \text{sterile}_{\mathcal{G}}(S)$ ; the set of recursive heads is denoted by  $\mathcal{H}(\mathcal{G})$ .<sup>1</sup>

The *tail* associated with  $H$  (and  $S$ ) is  $\text{pa}_{\mathcal{G}}(S)$ .

Intrinsic sets are central to the nested Markov property, as they are the sets parameterized by the kernels  $r_i$  in Definition 3.3.

**Example 4.2.** For the graph  $\mathcal{L}$  in Figure 2,  $\{2, 4, 6\}$  and  $\{3\}$  are districts and therefore intrinsic sets. The graph  $\mathcal{L}[\{2, 4, 6\}]$  is shown in Figure 4(a); applying  $\mathbf{m}$  appropriately to random-ancestral sets yields all the other intrinsic sets:  $\{2, 6\}$ ,  $\{4, 6\}$ ,  $\{2\}$ ,  $\{4\}$  and  $\{6\}$ . Each recursive head is equal to the associated intrinsic set.

**Definition 4.3.** Let  $B \subseteq V$  be a set of random vertices in  $\mathcal{G}$ . Suppose we alternately marginalize vertices that are not ancestors of  $B$ , and remove those which are not in the same district as some element of  $B$ :

$$\mathcal{G} \mapsto \mathbf{m}_{\text{anc}(B)}(\mathcal{G}) \qquad \mathcal{G} \mapsto \mathbf{d}_{\text{dis}_{\mathcal{G}}(B)}(\mathcal{G}) \qquad (4)$$

If these two operations change anything at all then they reduce the size of the set of random vertices; consequently this operation will eventually reach some stable point,

---

<sup>1</sup>Note that the definition of a recursive head differs from the *head* used in Evans and Richardson (2014) for ADMGs. We will see in Example 4.11 that  $\{E, Y\}$  is a recursive head in the graph in Figure 1(b), but one can check that it is not a head in the Evans and Richardson (2014) sense.



Figure 5: (a) A (C)ADMG  $\mathcal{G}$  and (b)  $\mathcal{G}_1 \equiv \mathfrak{d}_{\text{dis}(Y)}(\mathcal{G})$ .

which is a graph with random vertices denoted by  $I_{\mathcal{G}}(B)$ . Note that at each step of (4) the random vertices in the resulting graph always include  $B$ , so  $B \subseteq I_{\mathcal{G}}(B)$ .

If  $B$  is bidirected-connected then so is  $I_{\mathcal{G}}(B)$ ; in this case we call  $I_{\mathcal{G}}(B)$  the *intrinsic closure* of  $B$ .

**Example 4.4.** Let  $\mathcal{G}$  be the graph in Figure 5(a) and consider the intrinsic closure of the bidirected-connected set  $\{Y\}$ . The graph  $\mathfrak{m}_{\text{an}(Y)}(\mathcal{G})$  is just  $\mathcal{G}$ , since everything is an ancestor of  $Y$ . However  $\mathcal{G}_1 \equiv \mathfrak{d}_{\text{dis}(Y)}(\mathcal{G})$  gives the graph  $\mathcal{G}[\{X, Y\}]$  shown in Figure 5(b) in which  $Z$  is fixed, but the edges are all unchanged. It then becomes clear that repeatedly applying  $\mathfrak{m}$  and  $\mathfrak{d}$  will not result in any further changes to the graph. Hence the intrinsic closure is just the set of random vertices in this graph:  $I_{\mathcal{G}}(\{Y\}) = \{X, Y\}$ .

**Lemma 4.5.** *Let  $S$  be an intrinsic set with recursive head  $H$  in a graph  $\mathcal{G}$ . Then for any  $H \subseteq A \subseteq S$  we have  $I_{\mathcal{G}}(A) = S$ .*

Note that this shows, as we would hope, that each intrinsic set is its own intrinsic closure.

*Proof.* By the definition of  $H$ , every vertex in  $S$  is either in  $H$  or is a parent of some other element of  $S$ . Since  $S$  is bidirected-connected, the operations  $\mathfrak{d}_A$ ,  $\mathfrak{m}_A$  therefore cannot remove any element of  $S$  without also having removed an element of  $H$ , but this is not allowed since  $H \subseteq A$ . Hence no element of  $S$  is ever removed, and  $I_{\mathcal{G}}(A) \supseteq S$ .

Suppose that  $I_{\mathcal{G}}(A) \supset S$  and so  $B \equiv I_{\mathcal{G}}(A) \setminus S$  is non-empty. Every element of  $B$  is an ancestor of some other entry in  $I_{\mathcal{G}}(A)$ . In addition, every element of  $I_{\mathcal{G}}(A)$  is connected to  $A \subseteq S$  by bidirected paths through  $I_{\mathcal{G}}(A)$ , so  $I_{\mathcal{G}}(A)$  is, like  $S$ , a bidirected-connected set. Then it is clear from the definition of an intrinsic set that we cannot remove any element of  $B$  via operations of the form  $\mathfrak{m}, \mathfrak{d}$  without first removing some element of  $A \subseteq S$ . If  $B$  is non-empty then this implies  $S$  is not reachable, which contradicts the fact that  $S$  is intrinsic.  $\square$

Note that a corollary of this result is that recursive heads are in one-to-one correspondence with intrinsic sets: two distinct intrinsic sets may not have the same recursive head.

**Proposition 4.6.** *If  $B$  is a bidirected-connected set with intrinsic closure  $I_{\mathcal{G}}(B)$ , then the recursive head  $H$  associated with the intrinsic set  $I_{\mathcal{G}}(B)$  satisfies  $H \subseteq B$ .*

*Proof.* We must show that any sterile element of  $I_{\mathcal{G}}(B)$  in the CADMG  $\mathcal{G}[I_{\mathcal{G}}(B)]$  is in  $B$ . Suppose  $v \in \text{sterile}_{\mathcal{G}}(I_{\mathcal{G}}(B))$ ; but  $v$  is an ancestor of  $B$  in  $\mathcal{G}[I_{\mathcal{G}}(B)]$ , which

means that there is a directed path through the vertices in  $I_G(B)$  to some element  $b \in B$ . Then either  $v = b$ , or  $v \rightarrow w$  for some  $w \in I_G(B)$ , in which case  $v$  is not in the sterile set.  $\square$

**Lemma 4.7.** *Every singleton  $\{v\}$  for  $v \in V$  is a recursive head.*

*Proof.* Take the intrinsic closure  $I_G(\{v\})$  of  $v$ . Every element of  $I_G(\{v\})$  other than  $v$  is a parent of some other element of  $I_G(\{v\})$  by definition; therefore  $\{v\}$  is the sterile set, and a recursive head.  $\square$

**Lemma 4.8.** *Let  $\mathcal{G}$  be a CADMG, and  $\mathcal{G}'$  be a CADMG with random vertices  $V'$ , reachable from  $\mathcal{G}$ . Then the intrinsic sets of  $\mathcal{G}'$  are precisely the intrinsic sets of  $\mathcal{G}$  that are contained in  $V'$ , and their associated heads and tails are the same.*

*Proof.* Since  $\mathcal{G}' = \mathcal{G}[V']$  is reachable from  $\mathcal{G}$ , any intrinsic set in  $\mathcal{G}'$  is also an intrinsic set in  $\mathcal{G}$ . For the converse, suppose that  $D \subseteq V'$  is an intrinsic set in  $\mathcal{G}$ . Take the intrinsic closure of  $D$  in  $\mathcal{G}'$ , say  $C$ ; if  $C = D$  then we are done.

Suppose not, so that  $C \setminus D$  is non-empty. This occurs precisely when  $C$  is bidirected-connected, and every entry in  $C \setminus D$  is an ancestor of some other entry in  $C$ . But if this is true in  $\mathcal{G}'$  then it must also be true in  $\mathcal{G}$ , which contains any edges that  $\mathcal{G}'$  does; thus the intrinsic closure of  $D$  in  $\mathcal{G}$  is a strict superset of  $D$ . This contradicts Lemma 4.5.

By Lemma 2.8 the heads and tails associated with each intrinsic set are unchanged, since the parent sets of each random vertex are preserved.  $\square$

**Corollary 4.9.** *Let  $\mathcal{G}$  be a CADMG containing random-ancestral sets  $A_1, A_2$ . If  $H \in \mathcal{H}(\mathcal{G}[A_1])$  and  $H \in \mathcal{H}(\mathcal{G}[A_2])$  then  $H \in \mathcal{H}(\mathcal{G}[A_1 \cap A_2])$ .*

*Proof.* If  $A_1$  and  $A_2$  are random-ancestral, then so is  $A_1 \cap A_2$ , so the graph  $\mathcal{G}[A_1 \cap A_2]$  is reachable from  $\mathcal{G}$ . The result follows from Lemma 4.8.  $\square$

## 4.1 Partitions

We follow the approach of Evans and Richardson (2014) by defining partitions of sets via appropriate collections of subsets. Define a partial ordering  $\prec$  on recursive heads by  $H_1 \prec H_2$  whenever  $I_G(H_1) \subset I_G(H_2)$ .

**Definition 4.10.** Define a function  $\Phi_G$  on sets of random vertices  $C \subseteq V$  that ‘picks out’ the set of  $\prec$ -maximal elements of  $\mathcal{H}$  that are subsets of  $C$ . That is,

$$\Phi_G(C) \equiv \{H \in \mathcal{H} \mid H \subseteq C \text{ and } H \not\prec H' \text{ for all other } H' \subseteq C\}.$$

Define

$$\psi_G(C) \equiv C \setminus \bigcup_{D \in \Phi_G(C)} D.$$

Then recursively define a partition function  $\llbracket \cdot \rrbracket_G$  on subsets of  $V$  by  $\llbracket \emptyset \rrbracket_G = \emptyset$ , and

$$\llbracket W \rrbracket_G \equiv \Phi_G(W) \cup \llbracket \psi_G(W) \rrbracket_G.$$

For full details, including a proof that this definition does indeed define a partition, see the Appendix.

**Example 4.11.** The heads of the graph in Figure 1(b) are  $\{X\}, \{E\}, \{M\}, \{Y\}, \{E, Y\}$ , and the ordering requires that  $\{E\}$  and  $\{Y\}$  precede  $\{E, Y\}$ . Hence, for example

$$\begin{aligned}\llbracket \{X, E, Y\} \rrbracket_{\mathcal{G}} &= \{\{X\}, \{E, Y\}\} \\ \llbracket \{M, Y\} \rrbracket_{\mathcal{G}} &= \{\{M\}, \{Y\}\}.\end{aligned}$$

The partition  $[\cdot]_{\mathcal{G}}$  in Evans and Richardson (2014) made use of ‘heads’ rather than ‘recursive heads’, and therefore the partition obtained differs from the one here. For example, applied to the same graph as above,

$$[\{X, E, Y\}]_{\mathcal{G}} = \{\{X\}, \{E\}, \{Y\}\}.$$

**Lemma 4.12.** *If  $\mathcal{G}' = \mathcal{G}[D]$  is reachable from  $\mathcal{G}$  then  $\llbracket C \rrbracket_{\mathcal{G}'} = \llbracket C \rrbracket_{\mathcal{G}}$  for every  $C \subseteq D$ .*

*Proof.* By Lemma 4.8, the intrinsic sets of  $\mathcal{G}' = \mathcal{G}[D]$  are precisely the intrinsic sets of  $\mathcal{G}$  that are subsets of  $D$ , with the same associated recursive heads. Hence the result follows from the definition of  $\prec$ .  $\square$

**Lemma 4.13.** *If  $\mathcal{G}$  is such that  $V = D_1 \dot{\cup} D_2$  for sets  $D_1, D_2$  not connected by bidirected edges, then*

$$\llbracket C \rrbracket_{\mathcal{G}} = \llbracket C \cap D_1 \rrbracket_{\mathcal{G}} \cup \llbracket C \cap D_2 \rrbracket_{\mathcal{G}}.$$

*Proof.* Since every intrinsic set (and therefore recursive head) is a subset of either  $D_1$  or  $D_2$ , the result follows from Proposition A.4 in the Appendix.  $\square$

## 5 Parameterization

We are now in a position to introduce the parameterization.

**Definition 5.1.** Let  $\mathcal{G}$  be a CADMG with random vertices  $V$  and fixed vertices  $W$ . We say that  $p_{V|W}$  is *parameterized according to  $\mathcal{G}$* , and write  $p_{V|W} \in \mathcal{M}_p(\mathcal{G})$ , if it can be written in the form:

$$p_{V|W}(x_V | x_W) = \sum_{O \subseteq C \subseteq V} (-1)^{|C \setminus O|} \prod_{H \in \llbracket C \rrbracket_{\mathcal{G}}} q_H(x_T) \quad (5)$$

where  $q_H(x_T) \in \mathbb{R}$  for each  $H \in \mathcal{H}$ ,  $x_T \in \mathfrak{X}_T$ , and  $O \equiv \{v \in V \mid x_v = 0\}$ .

It will be shown in Section 5.1 that if  $p_{V|W}$  is of the above form then  $q_H(x_T) \in [0, 1]$  for all  $H$  and  $x_T$ . In fact, if the graph is interpreted causally, then each  $q_H(x_T)$  is the same as  $p_{H|T}(x_H \mid \text{do}(x_T))$ . It is also worth remarking on some special cases: if  $\mathcal{G}$  is a DAG then each  $H$  is a singleton  $\{h\}$ , and (5) is just the familiar parameterization in terms of conditional probability tables using corner-point identifiability constraints:  $q_H(x_T) = p_{h|\text{pa}(h)}(0_h \mid x_{\text{pa}(h)})$ . If  $\mathcal{G}$  has only bidirected edges then  $T = \emptyset$ , and (5) reduces to the parameterization given in Drton and Richardson (2008).

We will show that distributions are parameterized according to  $\mathcal{G}$  precisely when they recursively factorize according to  $\mathcal{G}$ , so that in fact  $\mathcal{M}_{rf}(\mathcal{G}) = \mathcal{M}_p(\mathcal{G})$ . In particular, a distribution of the form (5) satisfies properties (i) and (ii) of the recursive factorization. The next lemma shows that we can factorize such a distribution into pieces corresponding to the districts in  $\mathcal{G}$ , as required by property (i).

**Lemma 5.2.** *Let  $\mathcal{G}$  be a CADMG with random vertices  $V = D_1 \dot{\cup} \dots \dot{\cup} D_l$ , such that for  $i \neq j$  each  $D_i, D_j$  are not joined by any bidirected edges in  $\mathcal{G}$ . Then for any  $O \subseteq V$ ,*

$$\sum_{O \subseteq C \subseteq V} (-1)^{|C \setminus O|} \prod_{H \in \llbracket C \rrbracket_{\mathcal{G}}} q_H(x_T) = \prod_{i=1}^l \sum_{O_i \subseteq C \subseteq D_i} (-1)^{|C \setminus O_i|} \prod_{H \in \llbracket C \rrbracket_{\mathcal{G}}} q_H(x_T)$$

where  $O_i = O \cap D_i$ .

*Proof.* We prove the result for  $l = 2$ , from which the general result follows by induction. From Lemma 4.13

$$\prod_{H \in \llbracket C \rrbracket_{\mathcal{G}}} q_H(x_T) = \prod_{H \in \llbracket C \cap D_1 \rrbracket_{\mathcal{G}}} q_H(x_T) \times \prod_{H \in \llbracket C \cap D_2 \rrbracket_{\mathcal{G}}} q_H(x_T).$$

In addition if  $C_i = C \cap D_i$ , then  $C \setminus O = (C_1 \setminus O_1) \cup (C_2 \setminus O_2)$  and this is the union of two disjoint sets, so  $|C \setminus O| = |C_1 \setminus O_1| + |C_2 \setminus O_2|$ . Hence

$$\begin{aligned} \sum_{O \subseteq C \subseteq V} (-1)^{|C \setminus O|} \prod_{H \in \llbracket C \rrbracket_{\mathcal{G}}} q_H(x_T) &= \sum_{O \subseteq C \subseteq D_1 \cup D_2} (-1)^{|C \setminus O|} \prod_{H \in \llbracket C \cap D_1 \rrbracket_{\mathcal{G}}} q_H(x_T) \prod_{H \in \llbracket C \cap D_2 \rrbracket_{\mathcal{G}}} q_H(x_T) \\ &= \sum_{O_1 \subseteq C_1 \subseteq D_1} (-1)^{|C_1 \setminus O_1|} \prod_{H \in \llbracket C_1 \rrbracket_{\mathcal{G}}} q_H(x_T) \\ &\quad \times \sum_{O_2 \subseteq C_2 \subseteq D_2} (-1)^{|C_2 \setminus O_2|} \prod_{H \in \llbracket C_2 \rrbracket_{\mathcal{G}}} q_H(x_T). \end{aligned}$$

□

We now move to the main result of the paper.

**Theorem 5.3.**  *$p_{V|W}$  recursively factorizes according to  $\mathcal{G}$  if and only if it is parameterized according to  $\mathcal{G}$ .*

*Proof.* Throughout the proof we will write the partitions of vertices in a CADMG as  $\llbracket \cdot \rrbracket_{\mathcal{G}}$  regardless of which graph we are dealing with; since all the graphs we consider are reached from  $\mathcal{G}$ , this is justified by Lemma 4.12.

We proceed by induction on the size of  $V$ . If  $V = \{v\}$  then recursive factorization is essentially trivial, so the condition holds for any distribution. On the other hand, parameterization entails

$$p_{v|W}(0_v | x_W) = q_v(x_{\text{pa}(v)}), \quad p_{v|W}(1_v | x_W) = 1 - q_v(x_{\text{pa}(v)}),$$

which follows from setting  $q_v(x_{\text{pa}(v)}) = p_{v|W}(0_v | x_W)$  and elementary laws of probability; hence it also holds for any distribution.

( $\Leftarrow$ ) Now consider a general  $V$  and suppose  $p_{V|W}$  is parameterized according to  $\mathcal{G}$ . If  $\mathcal{G}$  has multiple districts then, by Lemma 5.2, the kernel factorizes into pieces which, by the induction hypothesis, are parameterized according to  $\mathcal{G}[D_i]$ .



Otherwise take any  $a \in \text{sterile}_{\mathcal{G}}(V)$ , and fix  $x_{V \setminus W \setminus \{a\}} \in \mathfrak{X}_{V \setminus W \setminus \{a\}}$ ; let  $O = \{v \in V \setminus \{a\} \mid x_v = 0\}$ , so then

$$\begin{aligned} \sum_{x_a} p(x_V \mid x_W) &= p(x_{V \setminus a}, 0_a \mid x_W) + p(x_{V \setminus a}, 1_a \mid x_W) \\ &= \sum_{O \cup \{a\} \subseteq C \subseteq V} (-1)^{|C \setminus (O \cup \{a\})|} \prod_{H \in \llbracket C \rrbracket_{\mathcal{G}}} q_H(x_T) + \sum_{O \subseteq C \subseteq V} (-1)^{|C \setminus O|} \prod_{H \in \llbracket C \rrbracket_{\mathcal{G}}} q_H(x_T) \\ &= \sum_{O \subseteq C \subseteq V \setminus \{a\}} (-1)^{|C \setminus (O \cup \{a\})|} \prod_{H \in \llbracket C \rrbracket_{\mathcal{G}}} q_H(x_T), \end{aligned}$$

which, by the induction hypothesis, recursively factorizes according to  $\mathcal{G}[V \setminus \{a\}] = \mathfrak{m}_{V \setminus a}(\mathcal{G})$ . This obviously extends to any random-ancestral margin  $V \setminus B$ . Hence  $p_{V \mid W}$  obeys properties (i) and (ii) of recursive factorization, and therefore recursively factorizes according to  $\mathcal{G}$ .

( $\Rightarrow$ ) Conversely, suppose that  $p_{V \mid W}$  recursively factorizes according to  $\mathcal{G}$ . In this direction we will strengthen the induction hypothesis slightly and show that if  $p_{V \mid W}$  recursively factorizes according to  $\mathcal{G}$  then  $p_{V \mid W}$  is parameterized according to  $\mathcal{G}$ , and that for each parameter  $q_H(x_T)$ , either  $p_{T \setminus W \mid W}(x_{T \setminus W} \mid x_{T \cap W}, y_{W \setminus T}) > 0$  for some  $y_{W \setminus T}$  in which case  $q_H(x_T)$  is uniquely recoverable from  $p_{V \mid W}$ , or  $p_{T \setminus W \mid W}(x_{T \setminus W} \mid x_{T \cap W}, y_{W \setminus T}) = 0$  for all  $y_{W \setminus T}$ , in which case  $q_H(x_T)$  can take any value. For the base case with  $|V| = 1$  the result follows from the derivation above.

If  $\mathcal{G}$  has multiple districts then, by definition,  $p_{V \mid W}$  factorizes into pieces which themselves recursively factorize according to the districts  $\mathcal{G}[D_i]$ , and by the induction hypothesis each factor is parameterized according to  $\mathcal{G}[D_i]$ . Applying Lemma 5.2 it follows that  $p_{V \mid W}$  is parameterized according to  $\mathcal{G}$ , and the parameters do not overlap by Lemma ???.

For uniqueness of  $q_H(y_T)$ , note that this parameter appears in the expansion for  $p_{V \mid W}(x_V \mid x_W)$  only if  $x_T = y_T$ . If  $p_{T \setminus W \mid W}(x_{T \setminus W} \mid x_W) > 0$  then the factorization of  $p_{V \mid W}$  is unique for these values by Proposition 3.6, and each factor is also positive for those values of  $x_T$  by Lemma 3.8; thus  $q_H(x_T)$  is uniquely recoverable from that factor by the induction hypothesis.

If  $p_{T \setminus W \mid W}(x_{T \setminus W} \mid x_W) = 0$  then by Lemma 3.8 there is some  $t \in T \setminus W$  and  $x_{V \setminus T}$  such that every version of  $p_{t \mid \text{pre}_{<}(t), W}(x_t \mid x_{\text{pre}_{<}(t)}, x_W) = 0$ . We split into two cases: either  $t$  is in the same district as  $H$ , or not; let  $D_1$  be the district containing  $H$ , and the associated kernel  $r_1$ . If  $t$  is in  $D_1$  then it follows from Proposition 3.6 that  $r_1(x_{T \cap D_1} \mid x_{T \setminus D_1}) = 0$ , and so by the induction hypothesis applied to  $\mathcal{G}[D_1]$  we get that  $q_H(x_T)$  can take any value. Otherwise if  $t$  is in a different district (say  $D_2$ ) then it follows from Proposition 3.6 that  $r_2(x_{T \cap D_2} \mid x_{\text{pa}(D_2) \setminus D_2}) = 0$ ; then clearly the value of every other factor, including  $r_1$ , is irrelevant at those  $x_T$ .

Now suppose  $\mathcal{G}$  has a single district  $V$ ; it follows from the definitions that  $V$  is intrinsic with recursive head  $H^* = \text{sterile}_{\mathcal{G}}(V)$  and tail  $T^* = (V \cup W) \setminus H^*$ . For any vertex  $h \in H^*$  the set  $V \setminus \{h\}$  is random-ancestral, so the margin  $p_{V \setminus h \mid W}$  recursively factorizes with respect to  $\mathcal{G}[V \setminus \{h\}]$ , and therefore (by the induction hypothesis) is also parameterized according to  $\mathcal{G}[V \setminus \{h\}]$ . Every head  $H$  other than  $H^*$  is found in at least one random-ancestral margin  $V \setminus \{h\}$  of  $\mathcal{G}$ , so applying the induction hypothesis to  $\mathcal{G}[V \setminus \{h\}]$  we obtain either a well defined parameter, or determine that its value is irrelevant.

If two or more random-ancestral margins contain the head  $H$ , note that by Corollary 4.9 there is a ‘smallest’ such margin  $p_{\text{an}(H)\setminus W|W}$  containing  $H$ ; all other random-ancestral margins must agree with this margin, and therefore by the induction hypothesis they will agree either on a value for  $q_H(x_T)$  or agree that it is arbitrary. So for every random-ancestral set  $A \subset V$  the margin  $\mathcal{G}[A]$  is parameterized according to  $p_{A|W}$  and any parameters that two or more of these margins jointly use can be chosen to agree.

The only head not found in any reachable subgraph is  $H^*$ , so the only parameter yet to be defined is  $q_{H^*}(x_{T^*})$ . We define this to be any version of  $p_{H^*|T^*}(0_{H^*} | x_{T^*})$ ; this is well defined if  $p(x_{T^*\setminus W} | x_{T^*\cap W}) > 0$ , and arbitrary otherwise. Then

$$p_{V|W}(0_{H^*}, x_{V\setminus H^*} | x_W) = q_{H^*}(x_{T^*}) \cdot p(x_{V\setminus H^*} | x_W).$$

Since  $V \setminus H^*$  is a random-ancestral margin of  $\mathcal{G}$ , it follows that  $p(x_{V\setminus H^*} | x_W)$  is parameterized according to  $\mathcal{G}[V \setminus H^*]$ , and so

$$\begin{aligned} &= p(0_{H^*} | x_{V\setminus H^*}, x_W) \cdot \prod_{O \subseteq C \subseteq V \setminus H^*} (-1)^{|C \setminus O|} \prod_{H \in \llbracket C \rrbracket_{\mathcal{G}}} q_H(x_T) \\ &= \prod_{O \subseteq C \subseteq V} (-1)^{|C \setminus O|} \prod_{H \in \llbracket C \rrbracket_{\mathcal{G}}} q_H(x_T). \end{aligned}$$

This gives the required result if  $x_h = 0$  for all  $h \in H^*$ . On the other hand, if  $x_h = 1_h$  for some  $h \in H^*$  then using a second induction on the number of zeros in  $x_{H^*}$  we have

$$\begin{aligned} p(x_{V\setminus h}, 1_h | x_W) &= p(x_{V\setminus h} | x_W) - p(x_{V\setminus h}, 0_h | x_W) \\ &= \sum_{O \subseteq C \subseteq V \setminus \{h\}} (-1)^{|C \setminus O|} \prod_{H \in \llbracket C \rrbracket_{\mathcal{G}}} q_H(x_T) - \sum_{O \cup \{h\} \subseteq C \subseteq V} (-1)^{|C \setminus (O \cup \{h\})|} \prod_{H \in \llbracket C \rrbracket_{\mathcal{G}}} q_H(x_T) \\ &= \sum_{O \subseteq C \subseteq V \setminus \{h\}} (-1)^{|C \setminus O|} \prod_{H \in \llbracket C \rrbracket_{\mathcal{G}}} q_H(x_T) + \sum_{O \cup \{h\} \subseteq C \subseteq V} (-1)^{|C \setminus O|} \prod_{H \in \llbracket C \rrbracket_{\mathcal{G}}} q_H(x_T) \\ &= \sum_{O \subseteq C \subseteq V} (-1)^{|C \setminus O|} \prod_{H \in \llbracket C \rrbracket_{\mathcal{G}}} q_H(x_T). \end{aligned}$$

Hence every probability  $p_{V|W}(x_V | x_W)$  is of the required form.  $\square$

## 5.1 Model Smoothness

From the proof above it follows that some of the parameters  $q_H(x_T)$  are just (versions of) the ordinary conditional probabilities  $P(X_H = 0 | X_T = x_T)$ , and hence the alternating sum is similar to the Möbius form of the parameterization studied in Evans and Richardson (2014) in the context of ‘ordinary’ Markov models. However we have already seen that not all of the parameters can be interpreted this way; recall the example in Section 3 for Figure 1(b). In this case, as noted in Example 3.5,  $q_{EY}(x, m) = r_{EY}(0, 0 | x, m)$  is not an ordinary conditional probability, but if the graph is interpreted causally then it is the conditional probability of  $\{E = Y = 0\}$  after intervening to fix  $\{X = x, M = m\}$ :

$$\begin{aligned} q_{EY}(x, m) &= p_{E|X}(0 | x) \cdot p_{Y|XME}(0 | x, m, 0) \\ &= P(Y = E = 0 | \text{do}(X = x, M = m)). \end{aligned}$$

This result holds more generally.

**Theorem 5.4.** *If  $p_{V|W}$  is strictly positive and recursively factorizes according to some CADMG  $\mathcal{G}$ , then all the parameters  $q_H(x_T)$  are unique and can be smoothly recovered from  $p_{V|W}$ .*

*In addition, if the graph is interpreted causally then*

$$q_H(x_T) = P(X_H = 0_H \mid \text{do}(X_T = x_T)).$$

*Proof.* The first claim follows directly from the proof of Theorem 5.3; the fact that  $p_{V|W}$  is strictly positive ensures that each parameter is always uniquely defined rather than being arbitrary.

The second part follows from the algorithm in Tian and Pearl (2002a).  $\square$

We remark that if the distribution is not strictly positive then the parameters  $q_H(x_T)$  are uniquely defined if and only if  $p(x_{V \cap T} \mid x_{W \cap T}, y_{W \setminus T}) > 0$  for some  $y_{W \setminus T}$ . In the case that  $W = \emptyset$  and  $\mathcal{G}$  is an ADMG, this reduces to  $q_H(x_T)$  being uniquely defined if and only if  $p(x_T) > 0$ .

Recall that  $\mathfrak{X}_v$  is a finite discrete state-space for the random variable  $X_v$ . Let  $\tilde{\mathfrak{X}}_v$  be the same set with some arbitrary entry removed (so that  $|\tilde{\mathfrak{X}}_v| = |\mathfrak{X}_v| - 1$ ). Then for any set  $C$  let  $\tilde{\mathfrak{X}}_C \equiv \times_{v \in C} \tilde{\mathfrak{X}}_v$ .

**Corollary 5.5.** *The set of strictly positive distributions obeying the recursive factorization property with respect to a CADMG  $\mathcal{G}$  is a curved exponential family of dimension*

$$d(\mathcal{G}) = \sum_{H \in \mathcal{H}(\mathcal{G})} |\tilde{\mathfrak{X}}_H| \cdot |\mathfrak{X}_T|.$$

*Proof.* Theorem 5.4 shows that there is a smooth (infinitely differentiable) map from positive distributions obeying the recursive factorization to the model parameters; it is clear from the form of the parameterization that the map from parameters to the probabilities is also smooth. The result follows by the same argument as Theorem 6.5 of Evans and Richardson (2014).  $\square$

This result allows us to invoke standard statistical theory within this class of models. For example, if  $\mathcal{G}'$  is a subgraph of  $\mathcal{G}$ , then we can perform a hypothesis test of  $H_0 : p_{V|W} \in \mathcal{M}_{rf}(\mathcal{G}')$  versus  $H_1 : p_{V|W} \in \mathcal{M}_{rf}(\mathcal{G})$  by comparing the likelihood ratio statistic to a  $\chi_k^2$  distribution, where  $k = d(\mathcal{G}) - d(\mathcal{G}')$ .

## 6 Examples

The Wisconsin Longitudinal Study (Hauser et al., 1957–2012) is a panel study of over 10,000 people who graduated from Wisconsin High Schools in 1957. We consider males who, when asked in 1975, had either been drafted or had not served in the military at all; after removing missing data this left 1,676 respondents. We wish to know whether, after controlling for family income and education, being drafted had a significant effect on future earnings.

The variables measured were:

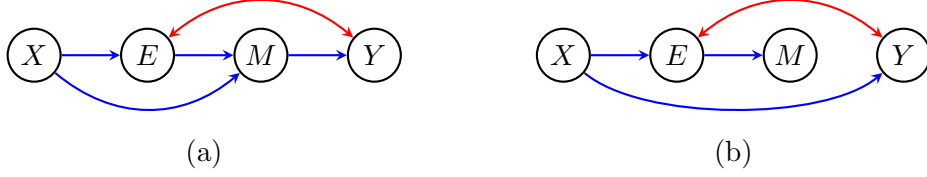


Figure 6: Two models for the Wisconsin military service data. (a) A proposed but rejected model; (b) a well-fitting model. See text for discussion.

- $X$ , an indicator of whether family income in 1957 was above \$5k;
- $Y$ , an indicator of whether the respondent’s income in 1992 was above \$37k;
- $M$ , indicator of whether the respondent was drafted into the military;
- $E$ , indicator of whether the respondent had education beyond high school.

Dichotomizations for  $X$ ,  $Y$  and  $E$  were chosen to be close to the median values of the original variables. The data are shown in Table 1; in each case the value 1 corresponds to the statement above being true, 0 otherwise. One possible model is that future income is unrelated to family income at the time of graduation after controlling for military service and level of education. This suggests the graph in Figure 6(a), where the directed edge from  $X$  to  $Y$  is not present. We can fit this model using the parameterization and an algorithm based on the one given by Evans and Richardson (2010); the resulting fit has a deviance of 31.3 on 2 degrees of freedom, strongly suggesting that the model should be rejected. Unsurprisingly, the graph in Figure 1(b) is also rejected for these data.

On the other hand the model shown in Figure 6(b) has a deviance of 5.57 on 6 degrees of freedom, which indicates a good fit. Note that this implies that there is no evidence of a significant effect of being drafted on future income, even though marginally there is a strong negative correlation. Models obtained by removing any additional edges are strongly rejected. Under this model the probability of having a high income in 1992 is estimated as 0.50 (standard error 0.018) if the family had high income, and 0.36 (0.016) if not.

In other words, we estimate

$$P(Y = 1 \mid \text{do}(X = 1)) = 0.50 \qquad P(Y = 1 \mid \text{do}(X = 0)) = 0.36,$$

indicating a strong causal effect.

### Acknowledgements

This research uses data from the Wisconsin Longitudinal Study (WLS) of the University of Wisconsin-Madison, which is supported principally by the National Institute on Aging.

### References

C. M. Bishop. *Pattern recognition and machine learning*. Springer, 2007.

$X = 0, E = 0$			$X = 1, E = 0$		
$M \backslash Y$	0	1	$M \backslash Y$	0	1
0	241	162	0	161	148
1	53	39	1	33	29

$X = 0, E = 1$			$X = 1, E = 1$		
$M \backslash Y$	0	1	$M \backslash Y$	0	1
0	82	176	0	113	364
1	13	16	1	16	30

Table 1: Data from the Wisconsin Longitudinal Study.

- A. Darwiche. *Modeling and reasoning with Bayesian networks*. Cambridge University Press, 2009.
- A. P. Dawid. Influence diagrams for causal modelling and inference. *International Statistical Review*, 70(2):161–189, 2002.
- M. Drton. Likelihood ratio tests and singularities. *Annals of Statistics*, pages 979–1012, 2009.
- M. Drton and T. S. Richardson. Binary models for marginal independence. *Journal of the Royal Statistical Society, Series B*, 70(2):287–309, 2008.
- R. J. Evans. Margins of discrete Bayesian networks. *arXiv:1501.02103*, 2015.
- R. J. Evans and T. S. Richardson. Maximum likelihood fitting of acyclic directed mixed graphs to binary data. In *Proceedings of the 26th conference on Uncertainty in Artificial Intelligence*, pages 177–184, 2010.
- R. J. Evans and T. S. Richardson. Marginal log-linear parameters for graphical Markov models. *Journal of the Royal Statistical Society, Series B*, 75(4):743–768, 2013.
- R. J. Evans and T. S. Richardson. Markovian acyclic directed mixed graphs for discrete data. *Annals of Statistics*, 42(4):1452–1482, 2014.
- R. M. Hauser, W. H. Sewell, and P. Herd. Wisconsin Longitudinal Study (WLS), 1957–2012. URL <http://www.ssc.wisc.edu/wlsresearch/documentation/>. Version 13.03, University of Wisconsin-Madison, WLS.
- J. C. Huang and B. C. Frey. Cumulative distribution networks and the derivative-sum-product algorithm. In *Proceedings of the 24th conference on Uncertainty in Artificial Intelligence*, pages 290–297, 2008.
- D. Mond, J. Smith, and D. van Straten. Stochastic factorizations, sandwiched simplices and the topology of the space of explanations. *Proceedings of the Royal Society A*, 459:2821–2845, 2003.
- J. Pearl. *Causality*. Cambridge University Press, Cambridge, UK, second edition, 2009.

- J. Pearl and T. S. Verma. A statistical semantics for causation. *Statistics and Computing*, 2(2):91–95, 1992.
- T. S. Richardson. Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30(1):145–157, 2003.
- J. M. Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9):1393–1512, 1986.
- I. Shpitser, R. J. Evans, T. S. Richardson, and J. M. Robins. Introduction to nested Markov models. *Behaviormetrika*, 41(1):3–39, 2014.
- R. Silva and Z. Ghahramani. The hidden life of latent variables: Bayesian learning with mixed graph models. *Journal of Machine Learning Research*, 10:1187–1238, 2009.
- R. Silva, C. Blundell, and Y. W. Teh. Mixed cumulative distribution networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 15, pages 670–678, 2011.
- J. Tian and J. Pearl. A general identification condition for causal effects. In *Proceedings of the 18th National Conference on Artificial Intelligence*. AAAI, 2002a.
- J. Tian and J. Pearl. On the testable implications of causal models with hidden variables. In *Proceedings of the Eighteenth conference on Uncertainty in Artificial Intelligence (UAI-02)*, pages 519–527. Morgan Kaufmann Publishers Inc., 2002b.
- T. S. Verma and J. Pearl. Equivalence and synthesis of causal models. In *Proceedings of the 7th Conference on Uncertainty in Artificial Intelligence (UAI-91)*, pages 255–268, 1991.
- N. Wermuth. Probability distributions with summary graph structure. *Bernoulli*, 17(3):845–879, 2011.

## A Partitions

Let  $V$  be an arbitrary finite set, and let  $\mathcal{H}$  be an arbitrary collection of non-empty subsets of  $V$ , with the restriction that  $\{v\} \in \mathcal{H}$  for all  $v \in V$  (i.e. all singletons are in  $\mathcal{H}$ ). A partial ordering  $\prec$  on the elements of  $\mathcal{H}$  will be said to be *partition suitable* if for any  $H_1, H_2 \in \mathcal{H}$  with  $H_1 \cap H_2 \neq \emptyset$ , there exists  $H^* \in \mathcal{H}$  such that  $H^* \subseteq H_1 \cup H_2$  and  $H_i \preceq H^*$  for each  $i = 1, 2$ . (Here  $H_1 \preceq H_2$  means  $H_1 \prec H_2$  or  $H_1 = H_2$ .)

Define a function  $\Phi$  on subsets of  $V$  such that  $\Phi(W)$  ‘picks out’ the set of  $\prec$ -maximal elements of  $\mathcal{H}$  that are subsets of  $W$ . That is,

$$\Phi(W) \equiv \{H \in \mathcal{H} \mid H \subseteq W \text{ and } H \not\prec H' \text{ for all other } H' \subseteq W\}.$$

Define

$$\psi(W) \equiv W \setminus \bigcup_{C \in \Phi(W)} C.$$

Then recursively define a partition function  $[\cdot]$  on subsets of  $V$  by  $[\emptyset] = \emptyset$ , and

$$[W] \equiv \Phi(W) \cup [\psi(W)].$$

It is clear that  $\cup_{A \in [W]} A = W$ .

The next proposition shows that  $[W]$  is indeed a partition of  $W$ .

**Proposition A.1.** *If  $H_1, H_2 \in \Phi(W)$  then  $H_1 \cap H_2 = \emptyset$ .*

*Proof.* Suppose  $H_1 \cap H_2 \neq \emptyset$ . Then by partition suitability, there exists  $H^* \subseteq H_1 \cup H_2$  with  $H^* \succeq H_1, H_2$ , and in particular  $H^* \succ H_i$  for at least one of  $i = 1, 2$ . Hence at least one of the  $H_i$  is not maximal in  $W$ .  $\square$

**Proposition A.2.** *If  $A \subseteq W_1 \subseteq W_2$ , and  $A \in \Phi(W_2)$  then  $A \in \Phi(W_1)$ .*

*Proof.* If  $A$  is a maximal head amongst the heads that are subsets of  $W_2$ , then it is certainly still maximal amongst heads that are subsets of  $W_1$ , since there are fewer such heads.  $\square$

**Proposition A.3.** *If  $C \in [W]$ , then  $[W] = \{C\} \cup [W \setminus C]$ .*

*Proof.* We proceed by induction on the size of  $W$ . If  $[W] = \{C\}$ , including any case in which  $|W| = 1$ , the result is trivial.

If  $C$  is not maximal with respect to  $\prec$  in  $W$ , then  $\Phi(W) = \Phi(W \setminus C)$ , and so

$$\begin{aligned} [W] &= \Phi(W) \cup [\psi(W)] \\ &= \Phi(W \setminus C) \cup [\psi(W)], \end{aligned}$$

and the problem reduces to showing that  $[\psi(W)] = \{C\} \cup [\psi(W \setminus C)]$ , which follows from the induction hypothesis. Thus, suppose  $C \in \Phi(W)$ .

Now by Proposition A.2,  $\Phi(W \setminus C) \cup \{C\} \supseteq \Phi(W)$ , and if equality holds we are done. Otherwise let  $C_1, \dots, C_k$  be the sets in  $\Phi(W \setminus C)$  but not in  $\Phi(W)$ . These sets are maximal in  $W \setminus C$ , so they are in  $\Phi(\psi(W))$  by Proposition A.2, since by hypothesis,  $\psi(W) \subseteq W \setminus C$ . Then the problem reduces to showing that

$$[\psi(W)] = \{C_1, \dots, C_k\} \cup [\psi(W) \setminus (C_1 \cup \dots \cup C_k)],$$

which follows from repeated application of the induction hypothesis.  $\square$

**Proposition A.4.** *Let  $D_1, \dots, D_k$  be a partition of  $V$ , and suppose that each  $H \in \mathcal{H}$  is contained within some  $D_i$ . Let  $\prec$  be a partition-suitable partial ordering. Then*

$$[W] = \bigcup_{i=1}^k [W \cap D_i].$$

*Proof.* We prove the case  $k = 2$ , from which the general result follows by repeated applications. If either  $W \cap D_1$  or  $W \cap D_2$  are empty, then the result is trivial. By definitions

$$[W] = \Phi(W) \cup [\psi(W)];$$

$\psi(W)$  is strictly smaller than  $W$ , so by the induction hypothesis

$$[W] = \Phi(W) \cup [\psi(W) \cap D_1] \cup [\psi(W) \cap D_2].$$

By hypothesis  $\Phi(W) = \mathcal{C}_1 \cup \mathcal{C}_2$  where each  $H \in \mathcal{C}_i$  is a subset of  $D_i$ ; since the elements of  $\mathcal{C}_i$  are maximal with respect to  $\prec$  in  $W$ , they are also maximal in  $W \cap D_i$ . Hence  $\mathcal{C}_i \subseteq \Phi(W \cap D_i)$ , and then applying Proposition A.3 gives

$$\mathcal{C}_i \cup [\psi(W) \cap D_i] = [W \cap D_i],$$

because  $(\psi(W) \cap D_i) \cup \bigcup \mathcal{C}_i = W \cap D_i$ . Hence the result.  $\square$

### A.1 Partition Suitability of Head Ordering

The next result, together with the previous one, shows that the partition defined in Section 4 for recursive heads is indeed a partition.

**Proposition A.5.**  *$\prec$  is partition suitable for  $\mathcal{H}(\mathcal{G})$ .*

*Proof.* Lemma 4.7 shows that  $\mathcal{H}$  contains the singleton vertices. Now suppose we have two recursive heads  $H_1, H_2$  with  $H_1 \cap H_2 \neq \emptyset$ . Let the associated intrinsic sets be  $S_1, S_2$ . Since  $S_1, S_2$  are bidirected connected sets and they share a common element,  $S_1 \cup S_2$  is also bidirected-connected. Let  $S^*$  be the intrinsic closure of  $S_1 \cup S_2$ , with recursive head  $H^*$ . Then  $S^*$  contains both  $S_1$  and  $S_2$ , and therefore  $H^* \succeq H_1, H_2$ .

By Proposition 4.6,  $H^* = \text{sterile}_{\mathcal{G}}(S^*) \subseteq S_1 \cup S_2$ ; By definition of a recursive head, any  $v \in S_1$  is either in  $H_1$  or is a parent of some other element of  $S_1$  (and the same for  $S_2$ ). Hence  $H^* \subseteq H_1 \cup H_2$ .  $\square$